Data Analysis Sample Case Study Report

Author:	Luis González
Email:	luisf.gonzalezv@yahoo.com
Date:	December 18, 2023

Table of Contents

Introduction

The following report describes a sample case study I (the author) completed as part of my training on Google´s Data Analytics Professional Certificate on Coursera. The case is based on free access real-world data from a bike share company in Chicago, called divvy. You can find out more about divvy by visiting their website here.

To use this company´s bike share service, users can either pay for a single ride, pay for a whole day of use, or purchase a yearly subscription (in general terms). Divvy offers open-access, historic monthly usage data of their services since April 2020. The datasets have been anonymized to protect user identities, differentiating only between casual users (that is, users who pay for single-use rides) and members. The purpose of this sample case study is to answer the hypothetical work questions of:

- How do casual users use bikes in comparison to members?

- Why would casual users buy annual memberships?

- How might advertising influence casual users to buy annual memberships?

Having understood the context of this exercise, we can move on to the data analysis methodology.

Methodology

The methodology process itself consists of several steps. Depending on the complexity and requirements of each data analysis/ data science problem, some steps of the process can be added or removed. In the present case, the objective is closest to developing a descriptive model, as we´re looking to find relationships within the data.

Note:

From step 3 onwards, the process is increasingly technical, and on this document I´ll be giving only the basic concept of the steps undertaken. If you wish to see the full explanation of how the data methodology took place, please refer to the document *methodology.ipynb*. It´s made on *jupyter notebooks*, and the code cells allow for a replication of the entire process, as well as an explanation for the reasoning made on each step.

1. Business Understanding

This part involves understanding the problem being dealt with. As explained in the introduction, we want to help a hypothetical bike company identify key behavioral differences between casual users and members, as well as provide advice on how best to advertise memberships to casual users.

2. Analytic Approach

We are now concerned with how data will help us solve the problem. It´s important to ensure the data available measures at least two different metrics by which we can compare casual users and members, in order to find differences and similarities between the two.

3. Data Requirements

The dataset is located in an online repository, downloadable in a compressed folder (.zip) format for historic monthly data, or quarterly in some cases. The repository is available here.

We want to work with the monthly data from 2022, the last complete year published when this case study was realized. Within the folder, we´ll find a .csv file with the dataset for that month.

4. Data Collection

The data for every month in 2022 was downloaded, and the .csv files moved into a common work folder with consistent naming. This was done manually, without the aid of code.

5. Data Understanding

It´s important to run some exploratory analysis on the datasets, to know what we´re working with and how best to tackle it. Using python *pandas* dataframes, it was discovered the whole dataset has over 5 million observations, each with 13 columns. The attributes of most interest to the study are the type of bike used (classic or electric), the spatial coordinates of the start and end of the trips (in latitude and longitude), and the start and end times of each ride, accurate to the last second.

6. Data Preparation

Given the size of the complete dataset is fairly extensive, it was decided to use an SQL database in conjunction with python to move and handle the data. SQL databases are optimized for very quick access and management of data. This was accomplished using python´s built-in *sqlite3* library.

After the data of all twelve files was moved to one sole location, the process of cleaning the dataset of invalid entries was commenced. New attributes were created, including trip duration in minutes and trip distance in km. For these last attribute, it was assumed most trips are commutes, and therefore in average users travel directly from their start location to their end one. Invalid entries, like those with locations outside Chicago, or with negative or very short time duration (less than one minute) were deleted from the database.

7. Finding Insights (Modeling)

With clean and accessible data, we can move on to finding relationships within the data. Queries in SQL were run to compare casual users and members on different attributes, like most used bike, number of trips, average trip duration and average trip distance. Also, to answer advertising-related questions, queries were run to identify the most popular commutes to and from registered stations, as well as the most visited stations. The results obtained are presented in the next section.

Results

Tableau was chosen as the tool to represent the findings of the case study, because of its appealing aesthetics, as well as user-friendly interface. The following represent the findings of the case study:

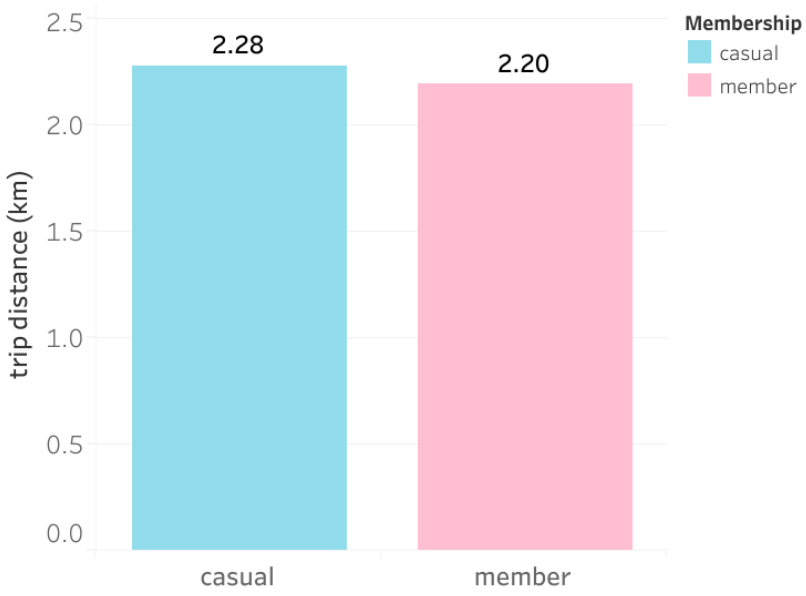

Figure 1: Average trip distance by membership type

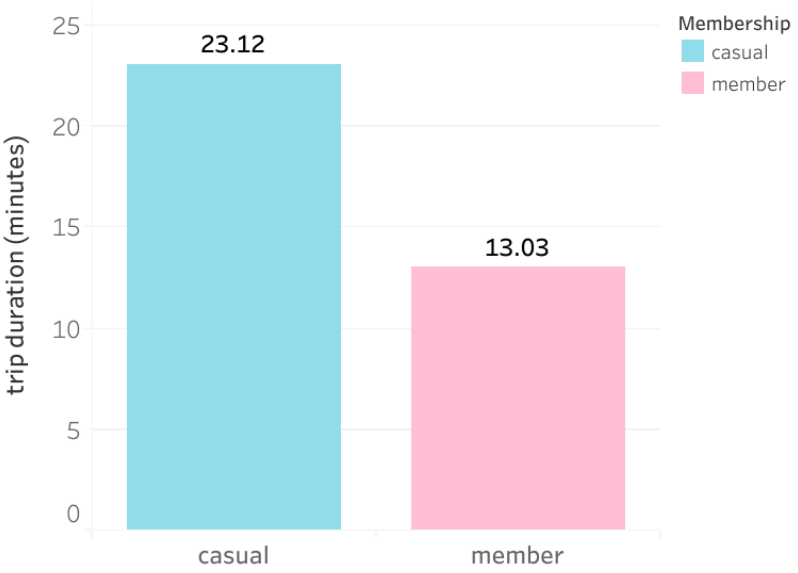

Figure 2: Average trip duration by membership type

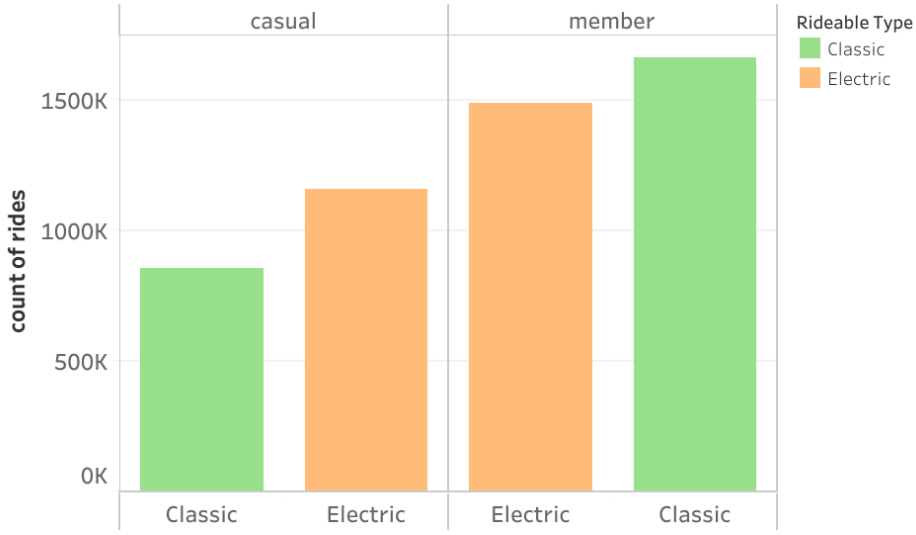

Figure 3: Count of rides by type of ride and membership

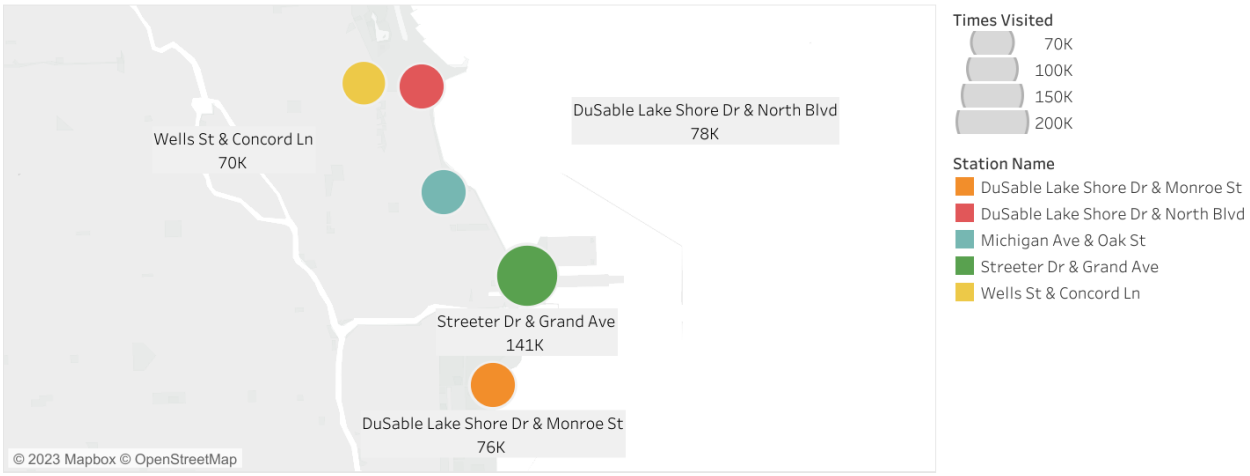

Figure 4: Most visited bike share stations in Chicago

Figure 4 has the added property of being a dynamic visual. If you wish, you´re welcome to visit the interactive version here, where you can modify the total visits filter, as well as the zoom of the map, to get a complete picture of the bike share activity in Chicago.

Discussion

The results show several interesting insights from the data. Let´s go through them, in the context of the questions presented in the introduction.

- **How do casual users use bikes in comparison to members?**

From figures 1 and 2, we can see that, on average, casual users travel similar distances to members, but they take almost twice as much time to complete their trips as members.

A possible explanation for this, is that members need more frequent use of the bikes for their day-to-day endeavors, therefore their trips are more likely to be previously planned, and take less time to complete. Casual users, on the other hand, may use the bikes on a more "ad hoc" basis, involving less planning and less frequent bike use, and take more time to complete their trips.

This is further reinforced by figure 3, where we can observe that electric bikes are more popular among casual users, whereas members slightly prefer to use classic bikes. Perhaps casual users reason they can get to their destination faster and easier on an electric bike, than on a classic one.

Lastly, we can see most trips are made by members (59.05 %), which means the company already does a decent job at convincing users to become full members.

- **Why would casual users buy annual memberships?**

For starters, casual users could buy annual memberships if they saw an economic benefit from doing so. Maybe the company could start friendly marketing campaigns, where they offer occasional free rides in exchange for the user´s email. They could then promote the economic benefits (if there are any) of becoming a member, for people who use their services past a certain weekly time threshold. Seeing how casual users on average take much lengthier rides than members, many would be likely to consider purchasing the subscription.

Another reason why casual users might look favorably on a membership is convenience. Having an annual membership saves the user of the burden of paying every time they use the service, and it gives them a slight comfort of mind knowing they can use the service whenever they needed. It might also help them become more organized and plan better their trips, so in average they save time as members do. All of the former are points that proper advertising could bring forward.

- **How might advertising influence casual users to buy annual memberships?**

In addition to the previous question, the company can target to use physical advertising on their most visited stations. According to figure 4, the 5 most popular bike share stations all reside in a relatively small area (in comparison to the size of the city), and that is the bay area. However, in this area more people are likely to already be members, so the benefits from advertising on these areas may be slim.

On the other hand, people residing on the metropolitan side of the city may not be so interested on the bike share service, or even know about it. Perhaps targeted publicity to these areas, as well as improvement in their stations´ infrastructure could help increase bike share activity in the other parts of the city, subsequently bringing in more customers and more members.

It´s important to clarify, however, the present findings are mostly inferential, given the limited information that can be extracted from the dataset. For a more detailed and reliable strategy on appropriate marketing strategy, more information is needed, like membership pricing plans and the total cost per ride of a casual user.

Conclusions

To conclude, the present case study looked to answer hypothetical business questions based on real world data of a bike share company based on Chicago. A combination of python and SQL allowed for easy management and study of a fairly extensive dataset (over 5 million records), and the main findings were conveyed in the most comprehensive and convincing way available to the audience using Tableau.

Despite the limitations of working with limited information, it was determined that casual users differ from members most significantly in that they take significantly longer time to complete their trips, and that they prefer to use electric bikes, most likely due to them only using the service for unexpected occasions.

It was also determined the most visited bike share stations all reside on the bay area of Chicago, implying that to increase the number of users, specifically members, the company can either attempt to increase their presence on other parts of the city, or advertise heavily on the bay area to influence more casual users to become members, leaning on the points mentioned in the discussion section.

It´s important to also mention the information extracted from the dataset is limited, and in a real world scenario more data is needed to device an appropriate marketing strategy.

Lastly, I thank you for your interest in this project :)